Modeling over-dispersion in binomial regression: a case study of Chronic Obstructive Pulmonary Disease patients through the SF-36 Health Survey

Josu Najera[1], *Dae-Jin Lee*[2], *Inmaculada Arostegui*[3]

[1]jnajera@bcamath.org and [2]dlee@bcamath.org, BCAM - Basque Center for Applied Mathematics

[3]inmaculada.arostegui@ehu.es, BCAM - Basque Center for Applied Mathematics and Department of Applied Mathematics, Statistics and Operations Research, UPV/EHU

In this work we present different approaches to deal with over-dispersion in Binomial outcomes in a regression context. We start considering a logistic model to show how the binomial distribution assumption do not perform well for over-dispersed simulated data. Indeed, a binomial distribution, $Y \sim \text{Bin}(n, p)$, where n is the number of trials and p the probability of success, is not adequate for this type of outcomes as the relationship between the expectation and the variance is not satisfied. The first and simplest approach consists of including a dispersion parameter to account for the over-dispersion such that $\text{E}(Y) = np$ and $\text{Var}(Y) = \phi np(1 - p)$. However, we show how this approach is only capable to account for very limited situations. To adequate more flexibility, we consider a second approach where the probability of success of the binomial distribution of each trial is not treated as fixed but random. Indeed, when this probability follows a beta distribution we have the beta-binomial model. The estimation of the models are done by maximum likelihood and iterative re-weighted least squares.

We consider outcomes from the Short Form-36 (SF-36) as a generic instrument to measure Health-Related Quality of Life (HRQoL) indicators of health status of patients with Chronic Obstructive Pulmonary Disease (COPD). Although these patient-reported outcomes are usually considered as continuous latent variables in the literature, in practice they are observed as ordinal. There is a proposal for recoding them to an ordinal form, based on the binomial distribution. We fitted a multivariate beta-binomial regression model where variable selection of demographic and clinical explanatory variables were performed. Finally we provided an interpretation of the estimated models for each of the scores of the SF-36 health survey of COPD patients in order to determine which factors are relevant on the health status of the patients in our study. An `R` package for the estimation of the models is under development by the authors.

Keywords: Binomial regression with over-dispersion, HRQoL, beta-binomial distribution.